第 58 卷 第 16 期/2021 年 8 月/激光与光电子学进展 研究论文

激光与光电子学进展 先进成像

基于改进 YOLOv3 的多目标跟踪算法研究

张相胜*，沈庆

江南大学物联网工程学院，轻工过程先进控制教育部重点实验室，江苏 无锡 214122

摘要 针对目前多目标跟踪过程中漏检率高和检测速率慢的问题，提出一种改进 YOLOv3 网络结构的多目标跟踪算法。首先，利用 K-means＋＋聚类算法对数据集中的目标边框进行聚类，根据聚类结果优化网络的先验框参数。然后，在 Darknet-53 特征提取层中引入深度可分离卷积模块，用深度可分离卷积代替标准卷积，减少参数量，并在 YOLO 预测层中引入 SENet 模块，利用 SENet 模块突出特征图的关键通道信息。最后，选定经典的 tracking-by-detection 框架，使用改进的 YOLOv3 算法来实现对目标信息的检测工作，跟踪部分选用 Deep-SORT 算法进行跟踪。实验结果表明，所提出的多目标跟踪算法能够有效地减小漏检率，同时兼顾了算法的检测精度和实时性。

关键词 图像处理；多目标跟踪；YOLOv3 网络；SENet 结构；深度可分离卷积；Deep-SORT 算法

中图分类号 TP391 **文献标志码** A **doi**：10.3788/LOP202158.1610004

Multitarget Tracking Algorithm Based on an Improved YOLOv3 Algorithm

Zhang Xiangsheng*, Shen Qing

School of Internet of Things Engineering, Key Laboratory of Advanced Control of Light Industry Process, Ministry of Education, Jiangnan University, Wuxi, Jiangsu 214122, *China*

Abstract To solve the problem of high missed rate and slow detection rate in the current multitarget tracking process, a multitarget tracking algorithm with an improved YOLOv3 network structure is proposed. First, the K-means＋＋ algorithm is utilized to cluster the target boundaries in the dataset. The priori parameters of the network are optimized using the clustering results. Then, the deep separable convolution module is employed instead of standard convolution in the Darknet-53 feature extraction layer, thereby reducing the number of parameters. In addition, the key channel information of the feature map is highlighted by applying the SENet module in the YOLO prediction layer. Finally, the improved YOLOv3 algorithm is used to implement the detection of a target in the classic tracking-by-detection framework. Meanwhile, the Deep-SORT algorithm is adopted in the tracking part. Experimental results show that the proposed multitarget tracking algorithm can effectively reduce the missed detection rate and take into account the detection accuracy and real-time performance, simultaneously.

Key words image processing; multi-target tracking; YOLOv3 network; SENet structure; deep separable convolution; Deep-SORT algorithm

OCIS codes 100.4996; 100.3008; 040.7290

1 引 言

近年来，随着深度学习在图像处理领域的快速发展，多目标跟踪已经逐渐成为了计算机视觉的研究热点，可用于视频监控、国防安全等领域[1-3]。跟踪方式可以分为检测跟踪和无检测跟踪两种，前者

收稿日期：2020-10-10；**修回日期**：2020-11-02；**录用日期**：2020-12-08

基金项目：国家自然科学基金(61773182)

通信作者：* zxs@jiangnan.edu.cn

需要检测到目标之后再进行跟踪，后者则需要手动初始化第一帧图片后再进行跟踪[4-5]。目标跟踪中，相继涌现出基于候选框的目标检测跟踪框架，如 R-CNN(Regions with CNN features)[6]、Faster R-CNN[7]，基于回归的目标检测跟踪框架，如 SSD(Single Shot Multibox Detector)[8]、YOLO(You Only Look Once)[9-11]系列。

基于回归框的检测网络是单步目标检测算法，不需要区域候选网络，直接利用网络即可产生相关的目标位置和类别信息，是一种端对端的检测网络，具有较快的检测速度。相对于 YOLO 算法和 YOLOv2 算法，YOLOv3 算法在准确度上有了很大的提升，同时提升了弱小目标的检测性能，在实际的检测跟踪中体现出较强的优势。在目标检测和跟踪的应用场合，干扰往往不可避免，即使采用较先进的 YOLOv3 网络，精确度依然较低。文献[12]通过对 YOLOv3 网络输出特征图进行上采样，并与残差块输出的特征图进行拼接，在 Darknet53 中增加残差单元，提高了检测目标的召回率、平均准确度，但算法的网络结构不够精简，实时性较差。

在多目标跟踪方面，文献[13]中使用匈牙利算法来解决数据关联问题，并对未遮挡目标和被遮挡目标分别采用不同的预测和更新方法，提高了目标跟踪的适应性，但当目标丢失超过一定帧数时，目标会丢失身份识别号。文献[14]结合 YOLOv3 与卡尔曼滤波器对行人目标进行跟踪，并采用质量评估网络改善目标遮挡的问题，但该算法对于相似目标干扰或者图像模糊情况的准确率较差。文献[15]提出的 SiamCNN 算法，是结合深度神经网络和上下文信息预测目标位置，该算法对遮挡问题没有相应的改善，漏检率较高。文献[16]提出了一种基于卷积神经网络(CNN)进行多目标跟踪的方法，使用一种新的 2D 在线环境分组策略，并利用卷积神经网络代替手工标注进行特征提取，具有较高的准确率和实时性。多目标跟踪 Deep-SORT[17](Simple Online and Realtime Tracking with a Deep Association Metricarxiv)算法相比之前的 SORT[18](Simple Online and Realtime Tracking)算法加入了深度表观特征，利用 CNN 在大规模数据集进行训练，并提取特征，增加了 CNN 对遗失和障碍的鲁棒性，使跟踪效果有明显的提升。

针对目前多目标跟踪过程中漏检率较高和实时性较差的问题，提出了一种改进 YOLOv3 网络的多目标跟踪算法。利用 K-means＋＋聚类算法代替 YOLOv3 网络的 K-means 聚类算法，对数据集中的边框进行聚类分析，根据聚类结果更新优化先验框参数。针对 YOLOv3 网络模型进行改进，使用深度可分离卷积[19]代替 YOLOv3 网络的标准卷积，减少计算量。把 SENet[20](Squeeze-and-Excitation Networks)模块嵌入到 YOLOv3 网络预测层中，可以增加整个网络对特征的选择和捕捉能力。最后把改进的 YOLOv3 网络模型与 Deep-SORT 算法相结合，完成多目标的跟踪。该方法在提高算法精度的同时，降低了运算量，提升了算法的运行速度。

2 YOLOv3 算法原理及改进

2.1 YOLOv3 算法

YOLOv3 算法将输入图像划分成 $S\times S$ 大小的网格，在每个网格内预测 B 个边界框，对 C 类目标进行检测，并输出每类目标的边界框和边界框的置信度。边界框的置信度定义为

$$R=P_{\mathrm{r}}\times\eta_{\mathrm{IOU}},\tag{1}$$

式中，P_{r} 为该边界框内存在对象的概率，η_{IOU} 为边界框与该对象实际边界框的交并比(IOU)。

通过设定阈值，将类别置信度低于阈值的边界框排除，随后边界框采用非极大值抑制(NMS)[21]方法进行筛选，得到边界框的 5 个参数为(x,y,w,h,p_c)，其中(x,y)为目标中心相对于单元格左上角的相对坐标，(w,h)是目标宽和高的坐标，p_c 代表目标第 c 类别的概率值。经过归一化处理以后，最终的网络输出为 $S\times S\times(5\times B+C)$。

2.2 YOLOv3 算法的网络参数优化

YOLOv3 算法中引入的 anchor(锚框)参数是一组宽高值固定的先验框，该先验框参数直接影响检测速度与精度。常用的方法是网络训练初始，通过 K-means 聚类方法得到先验框，其中 K-means 算法聚类中心的数量需要事先给定，且初始聚类中心也需人为确定。选取的聚类中心因具有较大的随机性，聚类结果不一样。利用 K-means＋＋算法随机性更小的特点，代替 K-means 算法对样本标签聚类进行分析。

为避免预测框和先验框边框之间产生更多误差，采用两者间交并比替代原始算法中的欧氏距离作为目标函数，目标函数定义为

$$D=\min\sum_{i=0}^{n}\sum_{j=0}^{k}[1-\eta_{\mathrm{IOU}}(i,j)],\tag{2}$$

式中，n 为样本数，k 为先验框数量。

2.3　YOLOv3 算法改进

2.3.1　深度可分离卷积原理

深度可分离卷积把通道和空间区域分开考虑，将标准卷积模块分解成两个分卷积模块。第一层为深度卷积，对每个输入通道使用单通道的轻量级滤波器；第二层为逐点卷积，即尺寸为 1 * 1 卷积，用来计算输入通道的线性组合。在保证结果一致的情况下，实现了通道和区域的分离，有效地减少了计算量，减小了模型尺寸，提高了检测网络的实时性。

假设输入的特征映射尺寸为 $D_F * D_F * M$，采用的标准卷积如图 1 所示，若使用尺寸为 $D_K * D_K * M * N$ 的卷积核进行卷积，输出的特征映射尺寸为 $D_G * D_G * M$。M 为输入的通道数，N 为输出通道数，所以标准卷积的计算量为：$D_K * D_K * M * N * D_F * D_F$。

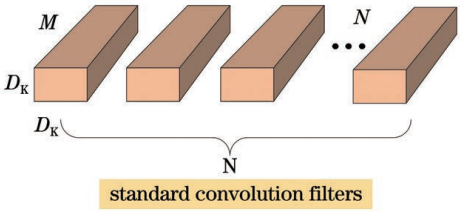

图 1　标准卷积结构

Fig. 1　Structure of standard convolution filters

如图2和图3所示，深度可分离卷积将标准卷积拆分为深度卷积和逐点卷积，深度卷积负责滤波，使用大小为 $D_K * D_K * 1 * N$ 的卷积核进行卷积，则计算量为 $D_K * D_K * M * D_F * D_F$。逐点卷积负责通道合并，使用大小为 $1 * 1 * M * N$ 的卷积核进行卷积，则计算量为 $M * N * D_F * D_F$。

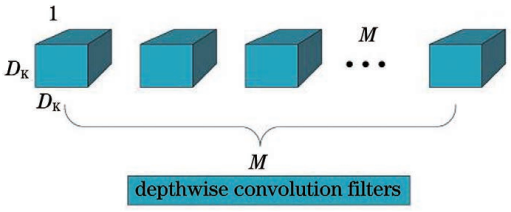

图 2　深度卷积结构

Fig. 2　Structure of depthwise convolution filters

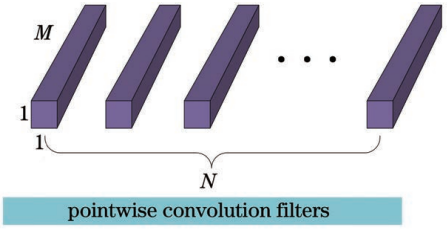

图 3　逐点卷积结构

Fig. 3　Structure of pointwise convolution filters

深度可分离卷积模块的计算量为深度卷积模块与逐点卷积模块之和，即 $D_K * D_K * M * D_F * D_F + M * N * D_F * D_F$。深度可分离卷积与标准卷积计算量之比为

$$\frac{D_K * D_K * M * D_F * D_F + M * N * D_F * D_F}{D_K * D_K * M * N * D_F * D_F} = \frac{1}{N} + \frac{1}{D_K^2}。\tag{3}$$

由此可以发现，利用深度可分离卷积可以相应地减少计算量。

2.3.2　SENet 模块

SENet 模块是由 Momenta 公司的 Hu 等[20]提出，SENet 模块的设计目的是在网络训练过程中增强用于分类的特征图，学习不同通道间的特征图中的权重信息，从而加快网络的训练速度，同时增加整个网络对特征的选择和捕捉能力。图 4 是 SENet 模块的结构图，图中 $\boldsymbol{U}_c$ 表示输入特征图，$\boldsymbol{S}_c$ 表示输出权重，c 为通道数，Global pooling 表示全局池化，FC 为全连接层，ReLU 表示线性修正单元，Sigmoid 表示 Sigmoid 激活函数，F_{sq}，F_{ex}，F_{scale} 分别表示对特征图进行压缩、特征提取以及张量拼接操作。

F_{sq}，F_{ex}，F_{scale} 的映射公式分别表示为

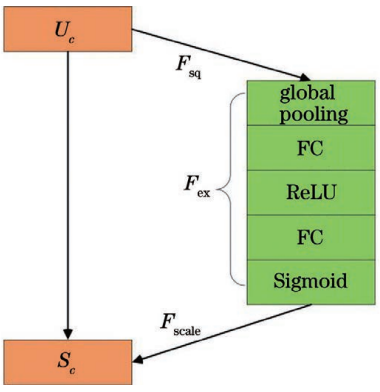

图 4　SENet 结构

Fig. 4　SENet structure

$$\boldsymbol{Z} = F_{sq}(\boldsymbol{U}_c) = \frac{1}{W \cdot H}\sum_{i=1}^{W}\sum_{j=1}^{H}\boldsymbol{U}_c(i,j),\tag{4}$$

$$\boldsymbol{S}_c = F_{ex}(\boldsymbol{Z},\boldsymbol{W}) = \sigma[\boldsymbol{g}(\boldsymbol{Z},\boldsymbol{W})] = \sigma[\boldsymbol{W}_2\boldsymbol{\delta}(\boldsymbol{W}_1\boldsymbol{Z})],\tag{5}$$

$$\tilde{X}=F_{\text{scale}}(\boldsymbol{U}_c,\boldsymbol{S}_c)=\boldsymbol{S}_c\cdot\boldsymbol{U}_c,\tag{6}$$

式中：W，H，C 分别代表输入特征图的宽、高、通道数；$\boldsymbol{Z}$ 表示产生的特征图压缩向量；$\boldsymbol{S}_c$ 表示 $\boldsymbol{U}_c$ 的第 C 个通道的特征图的权重；$\boldsymbol{W}_1$，$\boldsymbol{W}_2$ 表示两个全连接层中的权值；σ 表示 Sigmoid 激活函数；$\boldsymbol{\delta}$ 表示 ReLU 线性修正单元。

算法原理如下：1）对特征图压缩，将特征图转换成大小为 $1\times1\times C$ 的特征向量，代表该层 C 个特征图的全局数值分布情况，对应图 4 中 Global Pooling；2）第一个全连接层将特征降维到原来的 C/r'（r'为降维系数，是超参数，设定为 16），第二个全连接层通过升维恢复到原来的特征图大小，并使用 Sigmoid 函数对权重归一化；3）进行张量拼接，把权重 $\boldsymbol{S}_c$ 与原始特征图对应相乘得到最终输出。

针对 YOLOv3 网络处理特征时不能反映不同通道间特征的相关性和重要性的问题，嵌入 SENet 模块以加强网络的特征提取能力，降低漏检率。SENet 模块中包含两个全连接层，而全连接层的参数量相比其他网络层是最大的，因此为避免添加过量的 SENet 模块，影响算法的执行速度，分别在网络的三个预测层分支中嵌入 SENet 模块，以增加整个网络对特征的选择和捕捉能力。

2.3.3　YOLOv3 网络结构改进

YOLOv3 网络分为左边 Darknet-53 特征提取层和右边 YOLO 预测层，分别进行特征提取和多尺度预测。Darknet-53 由 5 个残差块（Resblock）构成，每个残差块由不同数量的卷积块循环一定数量之后融合构成。在 YOLO 预测层中，首先在 32 倍下采样、16 倍下采样和 8 倍下采样时使用 3 个不同尺度进行检测，然后融合上采样模块，用来提取深层特征，最后分别采用尺度大小为 13×13，26×26 和 52×52 的预测框来预测目标。

改进后的 YOLOv3 网络将 Darknet-53 特征提取层中的标准卷积替换为深度可分离卷积。交替使用深度卷积和逐点卷积，先利用 3×3 的深度卷积对输入图像的每个通道进行特征提取，再利用 1×1 的逐点卷积进行特征融合，缩减特征图尺寸的同时增加通道数。图 5 给出改进 YOLOv3 网络模型结构，图中 Dconv 模块表示卷积核大小为 3×3 的深度卷积，Pconv 代表卷积核大小为 1×1 的逐点卷积。在 YOLO 预测层每个分支中嵌入 SENet 模块。

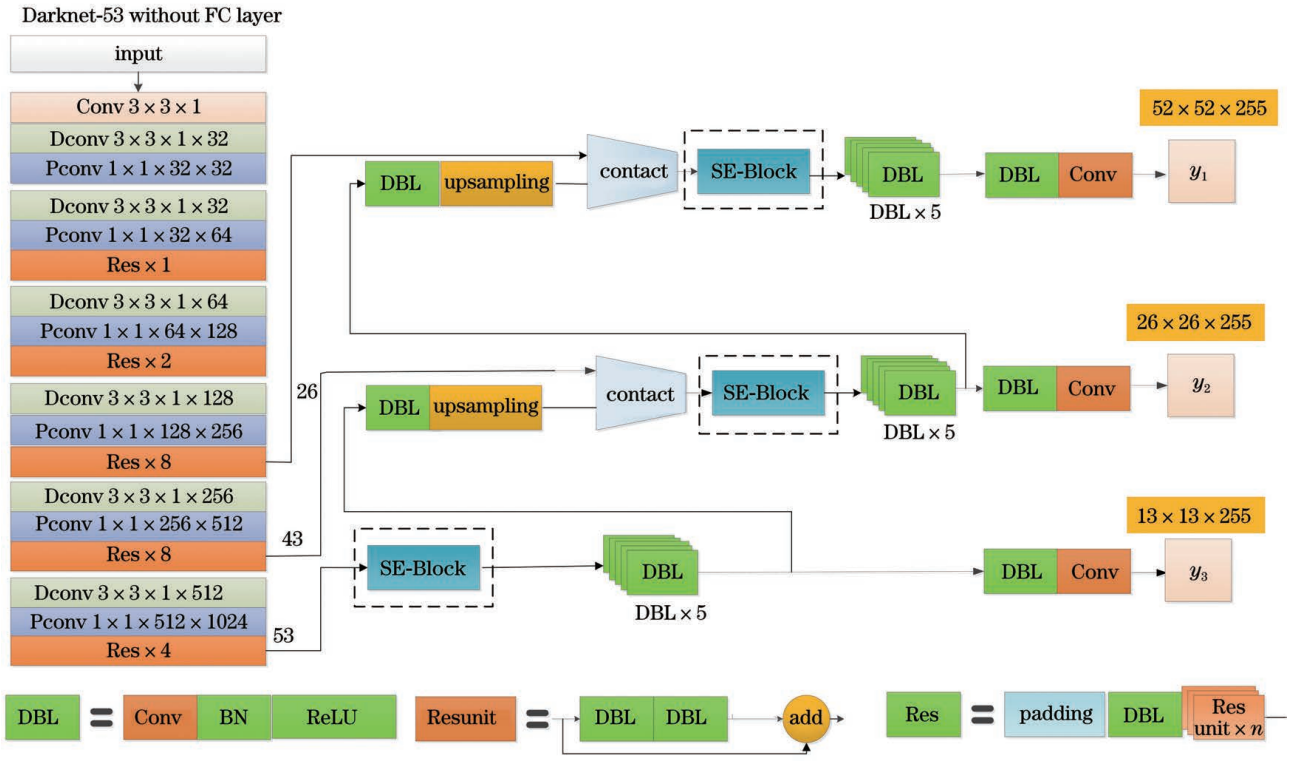

图 5　改进 YOLOv3 网络模型结构

Fig. 5　Improved YOLOv3 network model structure diagram

2.4　基于改进 YOLOv3 算法的多目标跟踪

本文使用改进的 YOLOv3 算法进行目标检测，与 Deep-SORT 算法相结合，进行多目标跟踪。算法流程如下。

1）目标检测：将输入的视频流进行目标检测，得到边框和特征信息，然后将经过检测后所得到的目标坐标信息(c_x,c_y,r,h,p)进行转换，得到 8 维向量$\boldsymbol{X}=[c_x,c_y,r,h,v_x,v_y,v_r,v_h]$，作为多目标跟踪算法的输入，其中，$p$ 为置信度分数，边界框的中心坐标为(c_x,c_y)，宽高比为 r，高为 h，v_x，v_y，v_r，v_h 代表 c_x，c_y，r，h 的速度变化值。

2）状态估计：首先使用卡尔曼滤波预测跟踪器在下一时刻的位置，然后基于检测结果来更新预测位置。

3）指派问题：利用匈牙利算法解决检测结果与跟踪预测结果的关联问题，同时考虑到运动信息的关联和目标外观信息的关联。

运动信息的关联，即采用已存在的卡尔曼滤波器预测结果和新检测结果之间的马氏距离来表达运动信息，马氏距离可表示为

$$d^{(1)}(i,j)=(\boldsymbol{d}_j-\boldsymbol{y}_i)^{\mathrm{T}}\boldsymbol{S}_i^{-1}(\boldsymbol{d}_j-\boldsymbol{y}_i),\quad(7)$$

式中，$d^{(1)}(i,j)$表示第 j 个检测框和第 i 条轨迹之间的运动匹配程度，$\boldsymbol{d}_j$ 表示第 j 个检测框的位置，$\boldsymbol{y}_i$ 表示第 i 条轨迹的状态向量，$\boldsymbol{S}_i$ 表示检测位置与平均位置之间的协方差矩阵。如果某次关联的马氏距离小于指定的阈值（阈值是从单独的训练集中得到的），则运动状态关联成功，指示函数的表达式为

$$b_{i,j}^{(1)}=\mathbb{I}[d^{(1)}(i,j)\leqslant\tau^{(1)}],\quad(8)$$

式中，$\mathbb{I}$ 为指示函数，设置阈值 $\tau^{(1)}=9.4847$。

在图像空间中使用卡尔曼滤波进行运动状态估计只是一个比较粗糙的预测，特别是相机存在运动时这种运动状态估计会使得马氏距离的关联方法失效，造成身份交换的误匹配。因此引入目标外观信息的关联方法，使用余弦距离来度量表观特征之间的距离，表达式为

$$d^{(2)}(i,j)=\min\{1-\boldsymbol{r}_i^{\mathrm{T}}\boldsymbol{r}_k^{(i)}\leqslant\boldsymbol{R}_i\},\quad(9)$$

式中，$\boldsymbol{r}$ 为经过 cosine 深度特征网络得到的特征向量，i 为轨迹数，限制条件为$\|\boldsymbol{r}_i\|=1$，$\boldsymbol{r}_k^{(i)}$ 为对应的跟踪目标的轨迹，$\boldsymbol{R}_i$ 用来存储最近 100 帧成功关联的特征向量。使用余弦距离来度量跟踪结果和检测结果对应的表观特征，实现了准确的预测。余弦部分的指示函数为

$$b_{i,j}^{(2)}=\mathbb{I}[d^{(2)}(i,j)\leqslant\tau^{(2)}],\quad(10)$$

式中，阈值 $\tau^{(2)}$ 一般被设为 0.2。

关联度量是通过将运动信息和外观信息加权来取得，即

$$c_{i,j}=\lambda d^{(1)}(i,j)+(1-\lambda)d^{(2)}(i,j),\quad(11)$$

式中，$c_{i,j}$ 表示综合匹配度，λ 为一个超参数，用来表达不同关联方法的权重，默认为 0。只有当 $c_{i,j}$ 位于两种度量阈值的交集内时，才认为实现了正确的关联。当指派完成后，将未匹配的检测结果和跟踪器筛选出来。

4）级联匹配和 IOU 匹配：当目标被长时间遮挡之后，卡尔曼滤波预测结果的正确性会降低，状态空间内的可观性也会相应降低，因此使用级联匹配对更加频繁出现的目标赋予更大的权重。对于未确认状态的跟踪器、未匹配的跟踪器和未匹配的检测，进行 IOU 匹配，再次使用匈牙利算法进行指派。

5）对于匹配的跟踪器进行参数更新，删除再次未匹配的跟踪器，初始化未匹配的检测结果为新目标。并判断视频流是否结束，若结束，退出循环；否则，进入下一帧检测。

相应的多目标跟踪算法整体流程如图 6 所示。

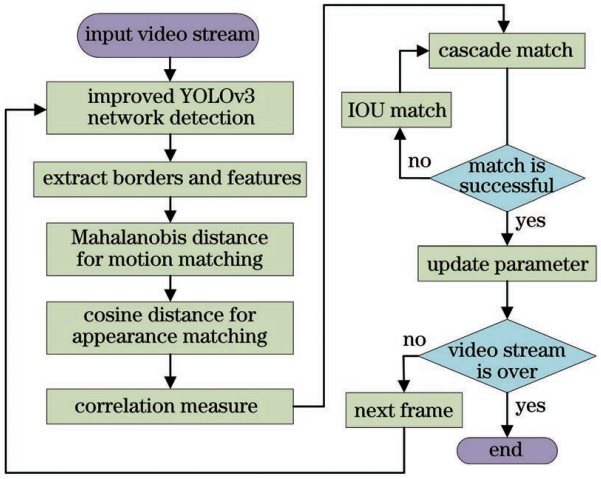

图 6　多目标跟踪算法流程

Fig. 6　Multi-target tracking algorithm flow

3　实验结果与分析

3.1　实验平台与数据集

算法利用 Keras 实现，在进行目标检测算法对比实验时，利用 VOC2007 数据集进行训练和测试；在进行多目标跟踪算法对比实验时，选用 MOT15，MOT16，ETHZ 等数据集进行测试。VOC2007 数据集中选取 10200 张图片为训练集，500 张为测试集，每张图片都有标注。训练平台环境为 Core(TM)i7-7700CPU@3.6 GHz，显卡是 Nvidia GeForce GTX 1080，运行内存为 8G。其他测试平台的 CPU 为 Inter(R)，Core(TM) i5-8250U，显卡是 Nvidia GeForce MX150，运行内存是 8G。软件平台是 Python3.6.3、Inter(R) win10 系统、CUDA8.0、

CUDNN6.0、Tensorflow-gpu1.4.0、Opencv3.0。

在训练过程中，为加快训练速度，防止过拟合，选用 SGD(stochastic gradient descent)作为基础迭代器，设置初始学习率为 0.001，采用 Adam 优化器，权重衰减系数为 0.005，迭代 12000 次。

3.2　先验框优化对比实验

使用改进的 YOLOv3 算法对选用数据集中的图片进行训练，分别用 K-means 聚类算法和 K-means++聚类算法对数据集标签进行聚类分析，选择不同数目的先验框，两种算法的平均交并比如图 7 所示。从图中可以看出，随着先验框数量不断增加，两种算法的平均交并比的趋势都在不断增大，但 K-means++聚类算法平均交并比的数值一直比 K-means 聚类算法更大，且趋势更为稳定，更能减小聚类偏差。同时由表1结果可以看出，在先验框的数目 k 值超过 9 时会出现大小较为相近的聚类结果，产生冗余。增加先验框的数量会导致模型的检测速度变慢，但综合考虑检测的准确性，本文最终选择 K-means++算法聚类生成的 9 个先验框，具体为(16,69)、(18,40)、(20,80)、(23,64)、(26,82)、(31,72)、(33,94)、(39,79)、(46,107)。

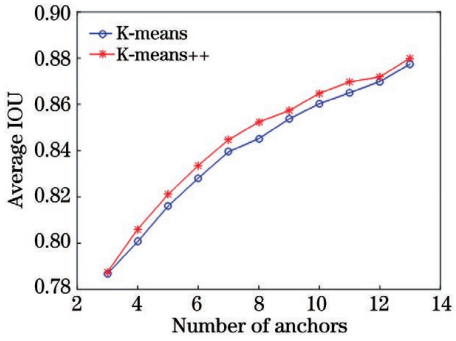

图 7　不同先验框数量的交并比

Fig. 7　Intersection-over-union of different number of anchors

表 1　不同先验框数量的先验框的尺寸

Table 1　Size of a priori boxes with different numbers of a priori boxes

k=7	k=8	k=9	k=10	k=11
(18,69)	(17,70)	(16,69)	**(18,40)**	(18,72)
(26,82)	(23,62)	(18,40)	**(18,75)**	(20,40)
(28,64)	(25,81)	(20,80)	(23,64)	(23,80)
(33,96)	(31,72)	(23,64)	(25,83)	(26,64)
(34,74)	(33,92)	(26,82)	(31,73)	(28,80)
(41,85)	(38,77)	(31,72)	(33,99)	**(34,89)**
(49,113)	(43,94)	(33,94)	(37,86)	**(34,72)**
	(52,122)	(39,79)	(38,74)	(36,105)
		(46,107)	(46,101)	(38,85)
			(62,132)	(44,83)
				(51,114)

3.3　目标检测算法对比实验

选取平均漏检率(Avgmisrate，ρ)、准确率(Precision，P_{re})、召回率(Recall，R_{re})、F_1 值和每秒检测帧数(FPS)作为评价指标来评估模型，其表达式分别为

$$\rho=\frac{X_{FN}}{X_{FN}+X_{TP}},\tag{12}$$

$$P_{re}=\frac{X_{TP}}{X_{TP}+X_{FP}},\tag{13}$$

$$R_{re}=\frac{X_{TP}}{X_{TP}+X_{FN}},\tag{14}$$

$$F_1=\frac{2\cdot P_{re}\cdot R_{re}}{P_{re}+R_{re}},\tag{15}$$

式中，X_{TP} 是真实正样本数量，X_{FP} 是虚假正样本数量，X_{FN} 是虚假负样本的数量。

调用训练得到的权重文件，选取 VOC2007 图片序列作为测试集进行测试，对比 Faster RCNN、YOLOv3 两种常用算法，使用相同数据集分别进行测试，检测结果如表 2 所示，从表中可以看出：由于加入了 SENet 模块，在平均漏检率方面，本文算法比 Faster RCNN 降低了 18.55 个百分点，比 YOLOv3 算法降低了 14.20 个百分点；在 F_1 值方面，本文算法相比 Faster RCNN 和 YOLOv3 提高并不明显，但是也稍有优势；在帧率(FPS)方面，由于本文算法把标准卷积替换为可分离的卷积，极大地减少了参数量，加快了网络检测的速度。

表 2　目标检测算法性能对比结果

Table 2　Target detection algorithm performance comparison results

Detection algorithm	Avgmisrate /%	F_1/%	FPS
Faster RCNN	32.15	88.57	5.52
YOLOv3	27.80	95.41	15.65
Our algorithm	**13.60**	**96.56**	**22.35**

3.4　多目标跟踪算法对比实验

选取 MOT15 数据集提供的 6 个序列对本文的算法进行测试，结果如表 3 所示。将所提算法与其他 7 种经典且当前性能较好的多目标跟踪算法进行

表 3　测试集在不同序列上的指标对比

Table 3　Comparison of the indicators of the test set on different sequences

Sequence	A_{MOT}/% ↑	P_{MOT}/% ↑	s_{ID} ↓	FN ↓
Venice-1	55.1	77.6	41	1212
KITTI-19	30.2	69.4	91	1626
KITTI-16	40.8	71.9	33	619
ETH-Crossing	66.3	80.4	15	252
PETS09-S2L2	55.6	73.1	177	2938
TUD-Crossing	76.8	72.8	21	202

比较，7 种多目标跟踪算法分别为 YOLOv3-SORT，Faster RCNN-Deep-SORT，YOLOv3-Deep-SORT，YOLOv3-Kalman[14]，SiamCNN[15]，MOTDT[22]，MDP[23]。结果如表 4 所示。各性能指标定义式为

$$A_{MOT}=1-\frac{N_{FN}+N_{FP}+s_{ID}}{N_{GT}}\in(-\infty,1], \tag{16}$$

$$P_{MOT}=\frac{\sum_{t,i}\eta_{t,i}}{\sum_{t}c_t}, \tag{17}$$

式中：A_{MOT} 表示多目标跟踪准确度，代表第一个性能指标；P_{MOT} 表示多目标跟踪精度，代表第二个性能指标；s_{ID} 表示目标身份交换的次数，代表第三个性能指标；N_{FN} 为第 4 个性能指标，代表整个视频漏报数量之和；N_{FP} 为整个视频误报数量之和；N_{GT} 是 Ground Truth 物体的数量；$\eta_{t,i}$ 表示第 t 帧的检测框与 N_{GT} 的交并比；c_t 表示第 t 帧的匹配个数。其中 A_{MOT}，P_{MOT} 越高表示效果越好；s_{ID} 和 N_{FN} 越低表示效果越好。

从表 4 可以看出，本文所提多目标跟踪算法在 A_{MOT}，P_{MOT}，s_{ID} 等性能指标上均以相应的优势领先其他算法，SiamCNN 算法结合深度神经网络和上下文信息来预测目标位置，但对遮挡问题没有相应的改善。MDP 算法将强化学习与深度学习相结合，利用马尔可夫决策方法提升了跟踪准确度，但没有结合目标当前状态进行分析，因此正确率有待提高。MOTDT 算法通过当前帧和过去帧的观测值来推断目标之后的状态，跟踪准确度有一定的提升，但在跟踪过程中无法校正目标位置，导致跟踪目标发生漂移。相比 YOLOv3-SORT，Faster RCNN-Deep-SORT，YOLOv3-Deep-SORT，YOLOv3-Kalman 这 4 种算法，本文算法在 YOLOv3 网络中，把标准卷积替换为可分离卷积，并且嵌入了 SENet 模块，同时把改进的 YOLOv3 网络与 Deep-SORT 算法相结合，因此在检测精度上表现良好，跟踪结果更精确。本文算法由于将标准卷积替换为可分离卷积，相比 YOLOv3-Deep-SORT 算法，减少了参数量，提升了 FPS 速度。

表 4　多目标跟踪算法评价指标对比

Table 4　Comparison of evaluation indexes of multi-target tracking algorithms

Algorithm	A_{MOT}/% ↑	P_{MOT}/% ↑	s_{ID} ↓	FPS ↑
YOLOv3-SORT	46.8	61.9	102	—
Faster RCNN-Deep-SORT	35.3	56.5	72	—
YOLOv3-Deep-SORT	54.8	68.0	68	2.8
YOLOv3-Kalman[14]	39.2	66.2	107	—
SiamCNN[15]	45.3	70.4	105	—
MOTDT[22]	57.3	75.3	70	—
MDP[23]	46.4	71.3	93	—
Our algorithm	56.0	**78.2**	**57**	**4.4**

3.5　算法对比效果分析

分别选取 MOT15，MOT16，ETHZ 多目标跟踪数据集中的序列进行多目标跟踪实验。

1）目标跟踪效果。图 8 为基于 MOT15 多目标跟踪数据集视频序列进行的多目标跟踪实验。本段视频为固定摄像头拍摄三岔路口行人的检测结果和流量统计情况。采用 YOLOv3-Deep-SORT 算法和本文方法在 MOT15 数据集 PETS09 序列上分别进行实验。对比同一帧图像可知，在视频第 68 帧和第 226 帧中，YOLOv3-Deep-SORT 算法把交通路障误识别为行人，本文算法能够对行人目标进行准确地检测与跟踪；在第 154 帧中，因路灯杆的遮挡较严重，YOLOv3-Deep-SORT 算法不能很好地框定目标，导致路灯杆后的行人漏检，而改进后的算法整体跟踪效果良好，同时能够有效地抑制因目标遮挡导致的漏检。

2）目标遮挡处理情况。图 9 为基于 MOT16 数据集的 MOT-06 序列进行的对比实验。其中第 103 帧图像中出现了大目标遮挡后面小目标的情况，YOLOv3-Deep-SORT 算法不能很好地跟踪目标，本文算法可以较好地完成跟踪任务；第 131 帧和第 140 帧图像中出现了目标同时被多个目标遮挡的情况，YOLOv3-Deep-SORT 在交错遮挡频繁的情况下容易产生跟踪误差，存在漏检率较高的问题，

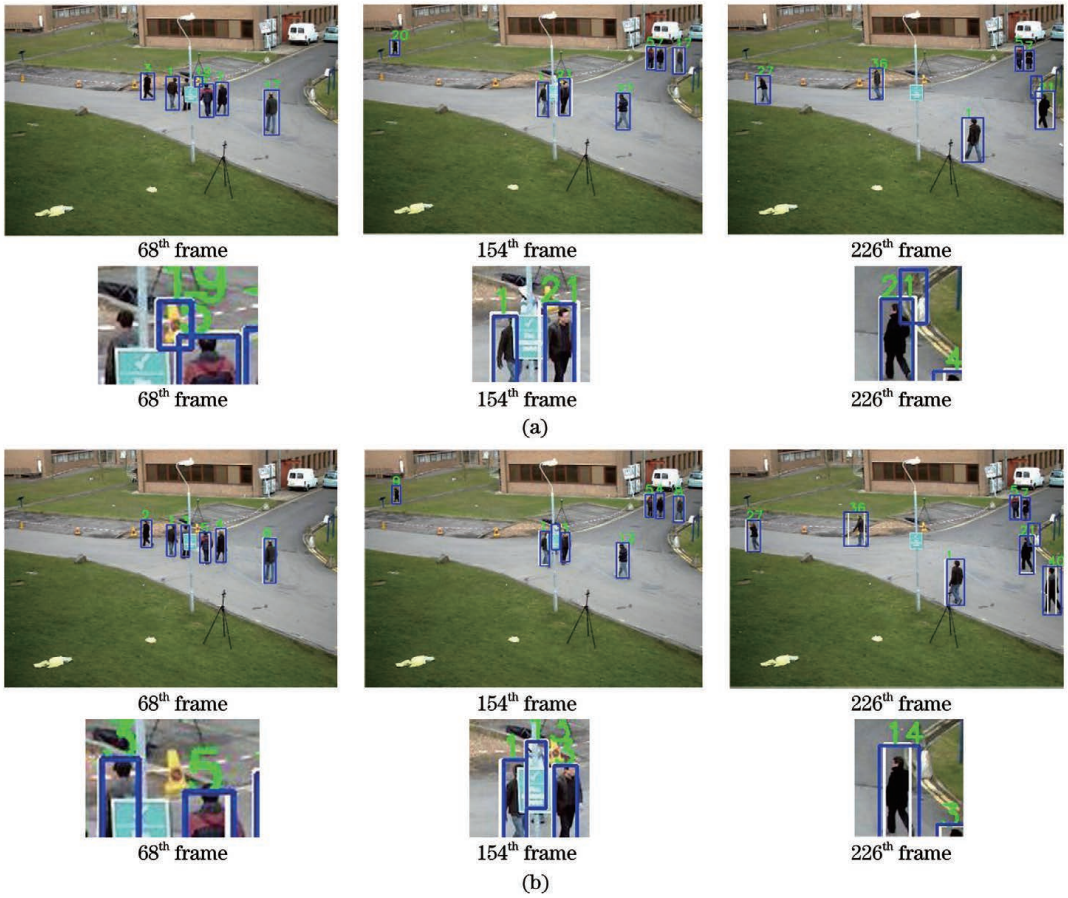

图 8 基于 MOT15-PETS09 序列的算法跟踪结果对比。(a) YOLOv3-Deep-SORT 跟踪结果;(b) 本文算法跟踪结果

Fig. 8 Comparison of algorithm tracking results based on MOT15-PETS09 sequence. (a) YOLOv3-Deep-SORT tracking results; (b) our algorithm tracking results

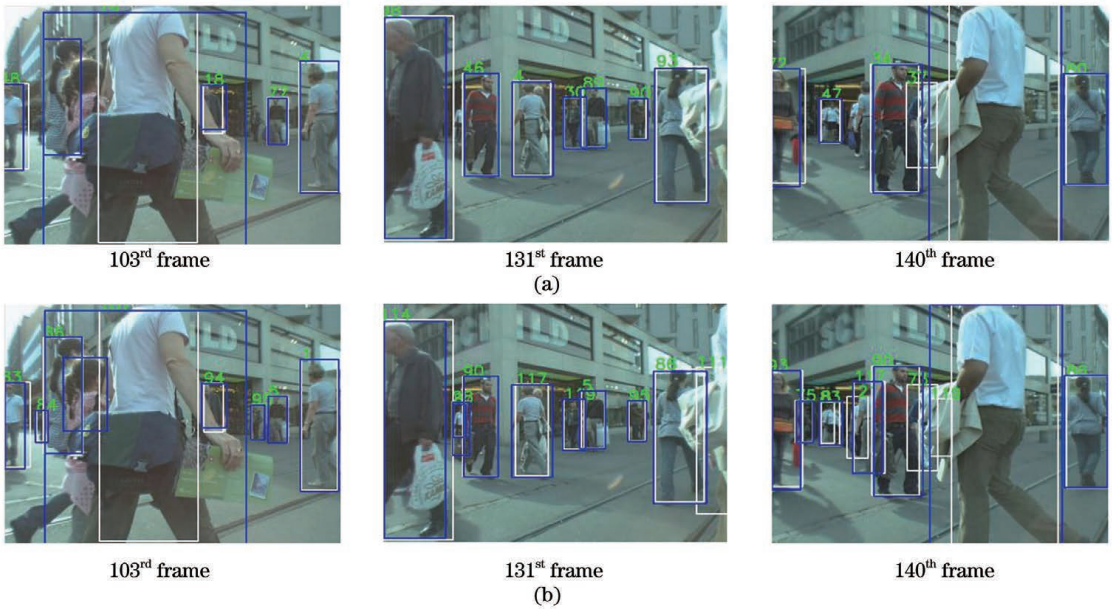

图 9 基于 MOT16-06 序列的算法跟踪结果对比。(a) YOLOv3-Deep-SORT 跟踪结果;(b) 本文算法跟踪结果

Fig. 9 Comparison of algorithm tracking results based on MOT16-06 sequence. (a) YOLOv3-Deep-SORT tracking results; (b) our algorithm tracking results

本文算法能够有效地持续跟踪被遮挡的目标。图 10 为具有小目标遮挡特征的 ETHZ 数据集 eth-02 序列实验。第 131 帧图像中出现了大量的小目标，本文算法可以有效地检测到小目标行人；第 209 帧和第 229 帧图像中出现了多个目标拥挤且目标被遮挡的情况，YOLOv3-Deep-SORT 算法则存在一定的漏检问题。本算法能够有效地持续跟踪被遮挡的目标。

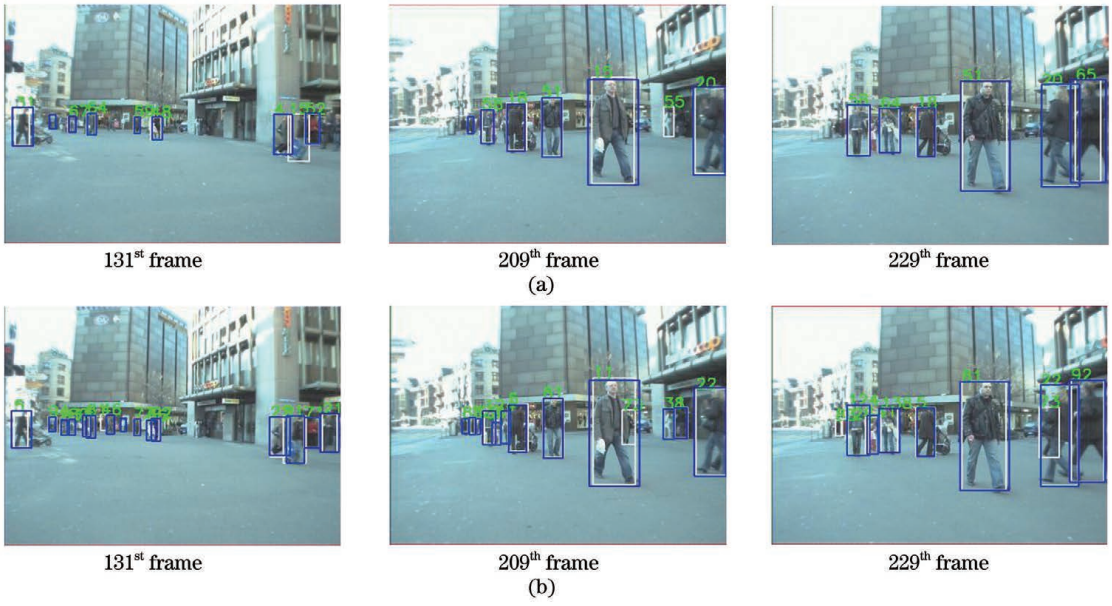

图 10　基于 ETHZ-eth02 序列的算法跟踪结果对比。(a) YOLOv3-Deep-SORT 跟踪结果；(b)本文算法跟踪结果

Fig. 10　Comparison of algorithm tracking results based on ETHZ-eth02 sequence. (a) YOLOv3-Deep-SORT tracking results; (b) our algorithm tracking results

4　结　　论

沿用 tracking-by-detection 框架，在 YOLOv3 网络和 Deep-SORT 算法的基础上，针对目标检测跟踪过程中的遮挡和漏检问题，用 K-means＋＋聚类方法优化先验框；利用深度可分离卷积模块替换 YOLOv3 网络中的标准卷积模块，进行特征提取；并将 SENet 模块嵌入到网络预测层中，与 Deep-SORT 多目标跟踪算法相结合。得到以下结论：

1) 使用 K-means 聚类＋＋方法优化先验框，比使用 K-means 聚类方法时的聚类误差更小。利用深度可分离卷积代替 YOLOv3 网络的标准卷积，加快了算法的运行速度。利用 SENet 模块反映出不同通道间特征的相关性和重要性，加强了网络的特征提取能力，提升了检测精度，并且结合 Deep-SORT 多目标跟踪算法，进而实现了整体算法快速且有效的跟踪。

2) 基于 MOT15，MOT16，ETHZ 多目标跟踪数据集对算法在多目标场景下的跟踪效果进行验证。结果表明，所提出的改进算法能够有效地避免目标遮挡导致的漏检情况，并且仍能保持较快的检测速度和较好的跟踪效果，具有较高的实时性。对实际的道路交通、视频监测等领域有一定的实用价值。

参　考　文　献

[1] Wang H Y, Wang L, Yin W R, et al. Multi-scale correlation filtering visual tracking algorithm combined with target detection [J]. Acta Optica Sinica, 2019, 39(1): 0115004.
王红雨, 汪梁, 尹午荣, 等. 结合目标检测的多尺度相关滤波视觉跟踪算法[J]. 光学学报, 2019, 39(1): 0115004.

[2] Liu M J, Cao Y Z, Zhu S Y, et al. Feature fusion video target tracking method based on convolutional neural network [J]. Laser & Optoelectronics Progress, 2020, 57(4): 041502.
刘美菊, 曹永战, 朱树云, 等. 基于卷积神经网络的特征融合视频目标跟踪方法[J]. 激光与光电子学进展, 2020, 57(4): 041502.

[3] Ju M R, Luo J N, Wang Z B, et al. Multi-scale target detection algorithm based on attention mechanism[J]. Acta Optica Sinica, 2020, 40(13): 1315002.
鞠默然, 罗江宁, 王仲博, 等. 融合注意力机制的多尺度目标检测算法[J]. 光学学报, 2020, 40(13): 1315002.

[4] Li X C, Liu X M, Cheng X N. A multi-target tracking algorithm based on YOLO detection[J]. Computer Engineering & Science, 2020, 42(4): 665-672.
李星辰, 柳晓鸣, 成晓男. 融合 YOLO 检测的多目标跟踪算法[J]. 计算机工程与科学, 2020, 42(4): 665-672.

[5] Li C Y, Yao J M, Lin Z X, et al. Object detection method based on improved YOLO lightweight network[J]. Laser & Optoelectronics Progress, 2020, 57(14): 141003.
李成跃, 姚剑敏, 林志贤, 等. 基于改进 YOLO 轻量化网络的目标检测方法[J]. 激光与光电子学进展, 2020, 57(14): 141003.

[6] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//2014 IEEE Conference on Computer Vision and Pattern Recognition, June 23-28, 2014, Columbus, OH, USA. New York: IEEE Press, 2014: 580-587.

[7] Ren S Q, He K M, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[8] Liu W, Anguelov D, Erhan D, et al. SSD: single shot MultiBox detector[M]//Leibe B, Matas J, Sebe N, et al. Computer vision-ECCV 2016. Lecture notes in computer science. Cham: Springer, 2016, 9905: 21-37.

[9] Redmon J, Divvala S, Girshick R, et al. You only look once: unified, real-time object detection[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 779-788.

[10] Redmon J, Farhadi A. YOLO9000: better, faster, stronger[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 6517-6525.

[11] Redmon J, Farhadi A. YOLOv3: an incremental improvement[EB/OL]. (2018-04-08)[2020-10-05]. https://arxiv.org/abs/1804.02767.

[12] Ju M R, Luo H B, Wang Z B, et al. Improved YOLOv3 algorithm and its application in small target detection[J]. Acta Optica Sinica, 2019, 39(7): 0715004.
鞠默然, 罗海波, 王仲博, 等. 改进的 YOLOv3 算法及其在小目标检测中的应用[J]. 光学学报, 2019, 39(7): 0715004.

[13] Yuan D L, Ji Q G. Multiple object tracking algorithm via collaborative motion status estimation[J]. Computer Science, 2017, 44(S2): 154-159.
袁大龙, 纪庆革. 协同运动状态估计的多目标跟踪算法[J]. 计算机科学, 2017, 44(S2): 154-159.

[14] Ren J M, Gong N S, Han Z Y. Multi-target tracking algorithm based on YOLOv3 and Kalman filter[J]. Computer Applications and Software, 2020, 37(5): 169-176.
任珈民, 宫宁生, 韩镇阳. 基于 YOLOv3 与卡尔曼滤波的多目标跟踪算法[J]. 计算机应用与软件, 2020, 37(5): 169-176.

[15] Leal-Taixé L, Canton-Ferrer C, Schindler K. Learning by tracking: siamese CNN for robust target association[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), June 26-July 1, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 418-425.

[16] Mahmoudi N, Ahadi S M, Rahmati M. Multi-target tracking using CNN-based features: CNNMTT[J]. Multimedia Tools and Applications, 2019, 78(6): 7077-7096.

[17] Wojke N, Bewley A, Paulus D. Simple online and realtime tracking with a deep association metric[C]// 2017 IEEE International Conference on Image Processing (ICIP), September 17-20, 2017, Beijing, China. New York: IEEE Press, 2017: 3645-3649.

[18] Bewley A, Ge Z Y, Ott L, et al. Simple online and realtime tracking[C]// 2016 IEEE International Conference on Image Processing (ICIP), September 25-28, 2016, Phoenix, AZ, USA. New York: IEEE Press, 2016: 3464-3468.

[19] Ji X S, Teng B. Detection of abnormal escalator behavior based on deep neural network[J]. Laser & Optoelectronics Progress, 2020, 57(6): 061010.
吉训生, 滕彬. 基于深度神经网络的扶梯异常行为检测[J]. 激光与光电子学进展, 2020, 57(6): 061010.

[20] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 7132-7141.

[21] Neubeck A, van Gool L. Efficient non-maximum suppression[C]//18th International Conference on Pattern Recognition (ICPR'06), August 20-24, 2006, Hong Kong, China. New York: IEEE Press, 2006: 850-855.

[22] Chen L, Ai H Z, Zhuang Z J, et al. Real-time multiple people tracking with deeply learned candidate selection and person re-identification[C]//2018 IEEE

International Conference on Multimedia and Expo (ICME), July 23-27, 2018, San Diego, CA, USA. New York: IEEE Press, 2018: 1-6.

[23] Xiang Y, Alahi A, Savarese S. Learning to track: online multi-object tracking by decision making[C]// 2015 IEEE International Conference on Computer Vision (ICCV), December 7-13, 2015, Santiago, Chile. New York: IEEE Press, 2015: 4705-4713.